Exploiting variable precision in GMRES

Gratton, Serge; Simon, Ehouarn; Titley-Peloquin, David; Toint, Philippe

Publication date:
2019

Document Version
Early version, also known as pre-print

Exploiting variable precision in GMRES

Serge Gratton* Ehouarn Simon† David Titley-Peloquin‡

Philippe Toint§

July 24, 2019

Abstract

We describe how variable precision floating point arithmetic can be used in the iterative solver GMRES. We show how the precision of the inner products carried out in the algorithm can be reduced as the iterations proceed, without affecting the convergence rate or final accuracy achieved by the iterates. Our analysis explicitly takes into account the resulting loss of orthogonality in the Arnoldi vectors. We also show how inexact matrix-vector products can be incorporated into this setting.

Keywords— variable precision arithmetic, inexact inner products, inexact matrix-vector products, Arnoldi algorithm, GMRES algorithm

AMS Subject Codes— 15A06, 65F10, 65F25, 97N20

1 Introduction

As highlighted in a recent SIAM News article [11], there is growing interest in the use of variable precision floating point arithmetic in numerical algorithms. In this paper, we describe how variable precision arithmetic can be exploited in the iterative solver GMRES. We show that the precision of some floating point operations carried out in the algorithm can be reduced as the iterations proceed, without affecting the convergence rate or final accuracy achieved by the iterates.

There is already a literature on the use of inexact matrix-vector products in GMRES and other Krylov subspace methods; see, e.g., [19, 6, 3, 7, 8] and the references therein. This work is not a simple extension of such results. To illustrate, when performing inexact matrix-vector products in GMRES, one obtains an inexact Arnoldi relation

$$AV_k + E_k = V_{k+1}H_k, \qquad V_k^T V_k = I. \tag{1}$$

*INPT-IRIT-ENSEEIHT, Toulouse, France (serge.gratton@toulouse-inp.fr).

†INPT-IRIT-ENSEEIHT, Toulouse, France (ehouarn.simon@toulouse-inp.fr). The work of this author was partially supported by the French National program LEFE (Les Enveloppes Fluides et l'Environnement).

‡McGill University, Montreal, Canada (david.titley-peloquin@mcgill.ca).

§The University of Namur, Namur, Belgium (philippe.toint@unamur.be).

On the other hand, if only inexact inner products are performed, the Arnoldi relation continues to hold exactly, but the orthogonality of the Arnoldi vectors is lost:

$$AV_k = V_{k+1}H_k, \qquad V_k^T V_k = I - F_k. \tag{2}$$

Thus, to understand the convergence behaviour and maximum attainable accuracy of GMRES implemented in variable precision arithmetic, it is absolutely necessary to understand the resulting loss of orthogonality in the Arnoldi vectors. We adapt techniques used in the rounding-error analysis of the Modified Gram-Schmidt (MGS) algorithm (see [1, 2] or [12] for a more recent survey) and of the MGS-GMRES algorithm (see [5, 9, 14]). We also introduce some new analysis techniques. For example, we show that (2) is equivalent to an exact Arnoldi relation in a non-standard inner product, and we analyze the convergence of GMRES with variable precision arithmetic in terms of exact GMRES in this inner product. For more results relating to GMRES in non-standard inner products, see, e.g., [10, 18] and the references therein.

We focus on inexact inner products and matrix-vector products (as opposed to the other saxpy operations involved in the algorithm) because these are the two most time-consuming operations in parallel computations. The rest of the paper is organized as follows. We start with a brief discussion of GMRES in non-standard inner products in Section 2. Next, in Section 3, we analyze GMRES with inexact inner products. We then show how inexact matrix-vector products can be incorporated into this setting in Section 4. Some numerical illustrations are presented in Section 5.

2 GMRES in weighted inner products

Shown below is the Arnoldi algorithm, with $\langle y, z\rangle = y^T z$ denoting the standard Euclidean inner product.

Algorithm 1 Arnoldi algorithm

Require: $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$

$\beta = \sqrt{\langle b, b\rangle}$
$v_1 = b/\beta$
for $j = 1, 2, \ldots$ **do**
  $w_j = Av_j$
  for $i = 1, \ldots, j$ **do**
    $h_{ij} = \langle v_i, w_j\rangle$
    $w_j = w_j - h_{ij}v_i$
  end for
  $h_{j+1,j} = \sqrt{\langle w_j, w_j\rangle}$
  $v_{j+1} = w_j/h_{j+1,j}$
end for

After k steps of the algorithm are performed in exact arithmetic, the output is $V_{k+1} = [v_1, \ldots, v_{k+1}] \in \mathbb{R}^{n\times(k+1)}$ and upper-Hessenberg $H_k \in \mathbb{R}^{(k+1)\times k}$ such that

$$v_1 = \frac{b}{\beta}, \qquad AV_k = V_{k+1}H_k, \qquad V_k^T V_k = I_k.$$

The columns of V_k form an orthonormal basis for the Krylov subspace $\mathcal{K}_k(A,b)$. In GMRES, we restrict x_k to this subspace: $x_k = V_k y_k$, where $y_k \in \mathbb{R}^k$ is the solution of

$$\min_y \|b - AV_k y\|_2 = \min_y \|V_{k+1}(\beta e_1 - H_k y)\|_2 = \min_y \|\beta e_1 - H_k y\|_2.$$

It follows that

$$\begin{aligned} x_k &= V_k y_k = V_k (H_k^T H_k)^{-1} H_k^T (\beta e_1) = V_k H_k^\dagger (\beta e_1), \\ r_k &= b - Ax_k = V_{k+1}(\beta e_1 - H_k y_k) = V_{k+1}(I - H_k H_k^\dagger)\beta e_1. \end{aligned} \tag{3}$$

Any given symmetric positive definite matrix W defines a weighted inner product $\langle y, z\rangle_W = y^T W z$ and associated norm $\|z\|_W = \sqrt{\langle z, z\rangle_W}$. Suppose we use this inner product instead of the standard Euclidean inner product in the Arnoldi algorithm. We use tildes to denote the resulting computed quantities. After k steps, the result is $\widetilde{V}_{k+1} = [\widetilde{v}_1, \ldots, \widetilde{v}_{k+1}]$ and upper-Hessenberg $\widetilde{H}_k \in \mathbb{R}^{(k+1)\times k}$ such that

$$\widetilde{v}_1 = \frac{b}{\|b\|_W} = \frac{b}{\widetilde{\beta}}, \qquad A\widetilde{V}_k = \widetilde{V}_{k+1}\widetilde{H}_k, \qquad \widetilde{V}_k^T W \widetilde{V}_k = I_k.$$

The columns of $\widetilde{V}_k$ form a W-orthonormal basis for $\mathcal{K}_k(A,b)$. Let $\widetilde{x}_k = \widetilde{V}_k \widetilde{y}_k$, where $\widetilde{y}_k \in \mathbb{R}^k$ is the solution of

$$\min_y \|b - A\widetilde{V}_k y\|_W = \min_y \|\widetilde{V}_{k+1}(\widetilde{\beta} e_1 - \widetilde{H}_k y)\|_W = \min_y \|\widetilde{\beta} e_1 - \widetilde{H}_k y\|_2,$$

so that

$$\widetilde{x}_k = \widetilde{V}_k \widetilde{H}_k^\dagger (\widetilde{\beta} e_1), \qquad \widetilde{r}_k = b - A\widetilde{x}_k = \widetilde{V}_{k+1}(I - \widetilde{H}_k \widetilde{H}_k^\dagger)\widetilde{\beta} e_1.$$

We denote the above algorithm W-GMRES.

The following lemma shows that if $\kappa_2(W)$ is small, the Euclidean norm of the residual vector in W-GMRES converges at essentially the same rate as in standard GMRES. The result is known; see e.g. [18]. We include a proof for completeness.

Lemma 1. *Let x_k and $\widetilde{x}_k$ denote the iterates computed by standard GMRES and W-GMRES, respectively, with corresponding residual vectors r_k and $\widetilde{r}_k$. Then*

$$1 \le \frac{\|\widetilde{r}_k\|_2}{\|r_k\|_2} \le \sqrt{\kappa_2(W)}. \tag{4}$$

Proof. Both x_k and $\widetilde{x}_k$ lie in the same Krylov subspace, $\mathcal{K}_k(A,b)$. Because x_k is chosen to minimize the Euclidean norm of the residual in $\mathcal{K}_k(A,b)$, while $\widetilde{x}_k$ minimizes the W-norm of the residual in $\mathcal{K}_k(A,b)$,

$$\|r_k\|_2 \le \|\widetilde{r}_k\|_2, \quad \|\widetilde{r}_k\|_W \le \|r_k\|_W.$$

Additionally, because for any vector z

$$\sigma_{\min}(W) \le \frac{z^T W z}{z^T z} = \frac{\|z\|_W^2}{\|z\|_2^2} \le \sigma_{\max}(W),$$

we have

$$\|r_k\|_2 \le \|\widetilde{r}_k\|_2 \le \frac{\|\widetilde{r}_k\|_W}{\sqrt{\sigma_{\min}(W)}} \le \frac{\|r_k\|_W}{\sqrt{\sigma_{\min}(W)}} \le \frac{\sqrt{\sigma_{\max}(W)}}{\sqrt{\sigma_{\min}(W)}}\|r_k\|_2,$$

from which (4) follows. ∎

3 GMRES with inexact inner products

3.1 Recovering orthogonality

We will show that the standard GMRES algorithm implemented with inexact inner products is equivalent to W-GMRES implemented exactly, for some well-conditioned matrix W. To this end, we need the following theorem.

Theorem 1. *Consider a given matrix $Q \in \mathbb{R}^{n\times k}$ of rank k such that*

$$Q^T Q = I_k - F. \tag{5}$$

If $\|F\|_2 \le \delta$ for some $\delta \in (0,1)$, then there exists a matrix M such that $I_n + M$ is symmetric positive definite and

$$Q^T(I_n + M)Q = I_k. \tag{6}$$

In other words, the columns of Q are exactly orthonormal in an inner product defined by $I_n + M$. Furthermore,

$$\kappa_2(I_n + M) \le \frac{1+\delta}{1-\delta}. \tag{7}$$

Proof. Note from (5) that the singular values of Q satisfy

$$\big(\sigma_i(Q)\big)^2 = \sigma_i(Q^T Q) = \sigma_i(I_k - F), \qquad i = 1,\dots,k.$$

Therefore,

$$\sqrt{1-\|F\|_2} \le \sigma_i(Q) \le \sqrt{1+\|F\|_2}, \qquad i = 1,\dots,k. \tag{8}$$

Equation (6) is equivalent to the linear matrix equation

$$Q^T M Q = I_k - Q^T Q.$$

It is straightforward to verify that one matrix M satisfying this equation is

$$\begin{aligned} M &= (Q^\dagger)^T (I_k - Q^T Q) Q^\dagger \\ &= Q(Q^T Q)^{-1}(I_k - Q^T Q)(Q^T Q)^{-1} Q^T. \end{aligned}$$

Notice that the above matrix M is symmetric. It can also be verified using the singular value decomposition of Q that the eigenvalues and singular values of $I_n + M$ are

$$\lambda_i(I_n + M) = \sigma_i(I_n + M) = \begin{cases} \big(\sigma_i(Q)\big)^{-2}, & i = 1,\dots,k, \\ 1, & i = k+1,\dots,n, \end{cases}$$

which implies that the matrix $I_n + M$ is positive definite. From the above and (8), provided $\|F\|_2 \le \delta < 1$,

$$\frac{1}{1+\delta} \le \frac{1}{\big(\sigma_{\max}(Q)\big)^2} \le \sigma_i(I_n + M) \le \frac{1}{\big(\sigma_{\min}(Q)\big)^2} \le \frac{1}{1-\delta}, \tag{9}$$

from which (7) follows. $\square$

Note that $\kappa_2(I_n + M)$ remains small even for values of δ close to 1. For example, suppose $\|I_k - Q^T Q\|_2 = \delta = 1/2$, indicating an extremely severe loss of orthogonality. Then $\kappa_2(I_n + M) \leq 3$, so Q still has exactly orthonormal columns in an inner product defined by a very well-conditioned matrix.

Remark 1. *Paige and his coauthors [2, 13, 17] have developed an alternative measure of loss of orthogonality. Given $Q \in \mathbb{R}^{n\times k}$ with normalized columns, the measure is $\|S\|_2$, where $S = (I+U)^{-1}U$ and U is the strictly upper-triangular part of $Q^T Q$. Additionally, orthogonality can be recovered by augmentation: the matrix $P = \left[\begin{smallmatrix} S \\ Q(I-S)\end{smallmatrix}\right]$ has orthonormal columns. This measure was used in the groundbreaking rounding error analysis of the MGS-GMRES algorithm [14]. In the present paper, under the condition $\|F\|_2 \leq \delta < 1$, we use the measure $\|F\|_2$ and recover orthogonality in the $(I + M)$ inner product. For future reference, Paige's approach is likely to be the most appropriate for analyzing the Lanczos and conjugate gradient algorithms, in which orthogonality is quickly lost and $\|F\|_2 > 1$ long before convergence.*

3.2 Bounding the loss of orthogonality

Suppose the inner products in the Arnoldi algorithm are computed inexactly, i.e., line 6 in Algorithm 1 is replaced by

$$h_{ij} = v_i^T w_j + \eta_{ij}, \tag{10}$$

with $|\eta_{ij}|$ bounded by some tolerance. We use tildes to denote the resulting computed quantities. It is straightforward to show that despite the inexact inner products in (10), the relation $AV_k = V_{k+1}H_k$ continues to hold exactly (under the assumption that all other operations besides the inner products are performed exactly). On the other hand, the orthogonality of the Arnoldi vectors is lost. We have

$$[b, AV_k] = V_{k+1}[\beta e_1, H_k], \qquad V_{k+1}^T V_{k+1} = I_{k+1} + F_k. \tag{11}$$

The relation between each η_{ij} and the overall loss of orthogonality F_k is very difficult to understand. To simplify the analysis we suppose that each v_j is normalized exactly. (This is not an uncommon assumption; see, e.g., [1] and [13].) Under this simplification, we have

$$F_k = \bar{U}_k + \bar{U}_k^T, \quad \bar{U}_k = \begin{bmatrix} 0_{k\times 1} & U_k \\ 0_{1\times 1} & 0_{1\times k}\end{bmatrix}, \quad U_k = \begin{bmatrix} v_1^T v_2 & \dots & v_1^T v_{k+1} \\ & \ddots & \vdots \\ & & v_k^T v_{k+1}\end{bmatrix}, \tag{12}$$

i.e., $U_k \in \mathbb{R}^{k\times k}$ contains the strictly upper-triangular part of F_k. Define

$$N_k = \begin{bmatrix} \eta_{11} & \dots & \eta_{1k} \\ & \ddots & \vdots \\ & & \eta_{kk}\end{bmatrix}, \quad R_k = \begin{bmatrix} h_{21} & \dots & h_{2k} \\ & \ddots & \vdots \\ & & h_{k+1,k}\end{bmatrix}. \tag{13}$$

Following Björck's seminal rounding error analysis of MGS [1], it can be shown that

$$N_k = -[0, U_k]H_k = -U_k R_k. \tag{14}$$

For completeness, a proof of (14) is provided in the appendix. Note that, assuming GMRES has not terminated by step k, i.e., $h_{j+1,j} \neq 0$ for $j = 1, \ldots, k$, then R_k must be invertible. Using (14), the following theorem shows how the convergence of GMRES with inexact inner products relates to that of exact GMRES. The idea is similar to [14, Section 5], in which the quantity $\|E_k R_k^{-1}\|_F$ must be bounded, where E_k is a matrix containing rounding errors.

Theorem 2. *Let $x_k^{(\mathrm{e})}$ denote the k-th iterate of standard GMRES, performed exactly, with residual vector $r_k^{(\mathrm{e})}$. Now suppose that the Arnoldi algorithm is run with inexact inner products as in* (10)*, so that* (11)–(14) *hold, and let x_k and r_k denote the resulting GMRES iterate and residual vector. If*

$$\|N_k R_k^{-1}\|_2 \leq \frac{\delta}{2} \tag{15}$$

for some $\delta \in (0, 1)$, then

$$1 \leq \frac{\|r_k\|_2}{\|r_k^{(\mathrm{e})}\|_2} \leq \sqrt{\frac{1+\delta}{1-\delta}}. \tag{16}$$

Proof. Consider the matrix F_k in (11). From (12) and (14), we have

$$\|F_k\|_2 \leq 2\|U_k\|_2 = 2\|N_k R_k^{-1}\|_2. \tag{17}$$

Thus, if (15) holds, $\|F_k\|_2 \leq \delta < 1$ and we can apply Theorem 1 with $Q = V_{k+1}$. There exists a symmetric positive definite matrix $W = I_n + M$ such that

$$[b, AV_k] = V_{k+1}[\beta e_1, H_k], \quad V_{k+1}^T W V_{k+1} = I_{k+1}, \quad \kappa_2(W) \leq \frac{1+\delta}{1-\delta}.$$

The Arnoldi algorithm implemented with inexact inner products has computed an W-orthonormal basis for $\mathcal{K}_k(A, b)$. The iterate x_k is the same as the iterate that would have been obtained by running W-GMRES exactly. The result follows from Lemma 1. $\square$

3.3 A strategy for bounding the η_{ij}

The challenge in applying Theorem 2 is bounding the tolerances η_{ij} at step j to ensure that (15) holds for all subsequent iterations k. Theorem 3 below leads to a practical strategy for bounding the η_{ij}. We will use

$$t_k = \beta e_1 - H_k y_k$$

to denote the residual computed in the GMRES subproblem at step k. We will use the known fact that for $j = 1, \ldots, k$,

$$|e_j^T y_k| \leq \|H_k^\dagger\|_2 \|t_{j-1}\|_2. \tag{18}$$

This follows from

$$e_j^T H_k^\dagger \underbrace{\begin{bmatrix} H_{j-1} & 0 \\ 0 & 0 \end{bmatrix}}_{\in \mathbb{R}^{(k+1)\times k}} \begin{bmatrix} y_{j-1} \\ 0_{k-j+1} \end{bmatrix} = e_j^T H_k^\dagger H_k \begin{bmatrix} y_{j-1} \\ 0 \end{bmatrix} = e_j^T \begin{bmatrix} y_{j-1} \\ 0 \end{bmatrix} = 0,$$

and thus

$$\begin{aligned}|e_j^T y_k| = |e_j^T H_k^\dagger \beta_1 e_1| &= \left| e_j^T H_k^\dagger \left(\beta_1 e_1 - \begin{bmatrix} H_{j-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{j-1} \\ 0 \end{bmatrix} \right) \right| \\ &= \left| e_j^T H_k^\dagger \begin{bmatrix} \beta_1 e_1 - H_{j-1} y_{j-1} \\ 0 \end{bmatrix} \right| \le \|H_k^\dagger\|_2 \|t_{j-1}\|_2.\end{aligned}$$

Additionally, in order to understand how $\|F_k\|_2$ increases as the residual norm decreases, we will need the following rather technical lemma, which is essentially a special case of [15, Theorem 4.1]. We defer its proof to the appendix.

Lemma 2. *Let y_k and t_k be the least squares solution and residual vector of*

$$\min_y \|\beta e_1 - H_k y\|_2.$$

Given $\epsilon > 0$, let D_k be any nonsingular matrix such that

$$\|D_k\|_2 \le \frac{\sigma_{\min}(H_k)\epsilon\|b\|_2}{\sqrt{2}\|t_k\|_2}. \tag{19}$$

Then

$$\frac{\|t_k\|_2}{\left(\epsilon^2\|b\|_2^2 + 2\|D_k y_k\|_2^2\right)^{1/2}} \le \sigma_{\min}\left(\left[\epsilon^{-1} e_1, H_k D_k^{-1}\right]\right) \le \frac{\|t_k\|_2}{\epsilon\|b\|_2}. \tag{20}$$

Theorem 3. *In the notation of Theorem 2 and Lemma 2, if for all steps $j = 1, \dots, k$ of GMRES all inner products are performed inexactly as in* (10) *with tolerances bounded by*

$$|\eta_{ij}| \le \eta_j \equiv \frac{\phi_j \epsilon \sigma_{\min}(H_k)}{\sqrt{2}} \frac{\|b\|_2}{\|t_{j-1}\|_2} \tag{21}$$

for any $\epsilon \in (0,1)$ and any positive numbers ϕ_j such that $\sum_{j=1}^k \phi_j^2 \le 1$, then at step k either (16) *holds with $\delta = 1/2$, or*

$$\frac{\|t_k\|_2}{\|b\|_2} \le 6k\epsilon, \tag{22}$$

implying that GMRES has converged to a relative residual of $6k\epsilon$.

Proof. If (21) holds, then in (13)

$$|N_k| \le \begin{bmatrix} \eta_1 & \eta_2 & \dots & \eta_k \\ & \eta_2 & \dots & \eta_k \\ & & \ddots & \vdots \\ & & & \eta_k \end{bmatrix} = E_k D_k,$$

where E_k is an upper-triangular matrix containing only ones in its upper-triangular part, so that $\|E_k\|_2 \le k$, and $D_k = \mathrm{diag}(\eta_1, \dots, \eta_k)$. Then in (15),

$$\begin{aligned}\|N_k R_k^{-1}\|_2 &\le \|N_k D_k^{-1}\|_2 \|D_k R_k^{-1}\|_2 \\ &\le \|E_k\|_2 \|D_k R_k^{-1}\|_2 \le k\|(R_k D_k^{-1})^{-1}\|_2.\end{aligned} \tag{23}$$

Let h_k^T denote the first row of H_k, so that $H_k = \begin{bmatrix} h_k^T \\ R_k \end{bmatrix}$. For any $\epsilon > 0$ we have

$$\begin{aligned}
\sigma_{\min}(R_k D_k^{-1}) &= \min_{\|u\|_2=\|v\|_2=1} u^T R_k D_k^{-1} v \\
&= \min_{\|u\|_2=\|v\|_2=1} [0, u^T] \begin{bmatrix} \epsilon^{-1} & h_k^T D_k^{-1} \\ 0 & R_k D_k^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ v \end{bmatrix} \\
&\geq \min_{\|u\|_2=\|v\|_2=1} u^T \begin{bmatrix} \epsilon^{-1} & h_k^T D_k^{-1} \\ 0 & R_k D_k^{-1} \end{bmatrix} v \\
&= \sigma_{\min}\left(\left[\epsilon^{-1} e_1, H_k D_k^{-1}\right]\right).
\end{aligned}$$

Therefore,

$$\|(R_k D_k^{-1})^{-1}\|_2 = \frac{1}{\sigma_{\min}(R_k D_k^{-1})} \leq \frac{1}{\sigma_{\min}\left(\left[\epsilon^{-1} e_1, H_k D_k^{-1}\right]\right)}.$$

Notice that if the η_j are chosen as in (21), D_k automatically satisfies (19). Using the lower bound in Lemma 2, and then (18) and (21), we obtain

$$\begin{aligned}
\|(R_k D_k^{-1})^{-1}\|_2 &\leq \frac{\left(\epsilon^2 \|b\|_2^2 + 2\|D_k y_k\|_2^2\right)^{1/2}}{\|t_k\|_2} \\
&= \frac{\left(\epsilon^2 \|b\|_2^2 + 2\sum_{j=1}^k \eta_j^2 (e_j^T y_k)^2\right)^{1/2}}{\|t_k\|_2} \\
&\leq \frac{\left(\epsilon^2 \|b\|_2^2 + \sum_{j=1}^k \phi_j^2 \epsilon^2 \|b\|_2^2\right)^{1/2}}{\|t_k\|_2} = \frac{\sqrt{2}\epsilon\|b\|_2}{\|t_k\|_2}.
\end{aligned}$$

Therefore, in (23),

$$\|N_k R_k^{-1}\|_2 \leq \frac{\sqrt{2}k\epsilon\|b\|_2}{\|t_k\|_2} \leq \frac{6k\epsilon\|b\|_2}{\|t_k\|_2}\frac{\delta}{2}$$

with $\delta = 1/2$. If (22) does not hold, then $\|N_k R_k^{-1}\|_2 \leq \delta/2$, which from Theorem 2 implies (16). Therefore, if the $|\eta_{ij}|$ are bounded by tolerances η_j chosen as in (21), either (16) holds with $\delta = 1/2$, or (22) holds. □

Theorem 3 can be interpreted as follows. If at all steps $j = 1, 2, \ldots$ of GMRES the inner products are computed inaccurately with tolerances η_j in (21), then convergence at the same rate as exact GMRES is achieved until a relative residual of essentially $k\epsilon$ is reached. Notice that η_j is inversely proportional to the residual norm. This allows the inner products to be computed more and more inaccurately as as the iterations proceed.

If no more than $K_{\max}$ iterations are to be performed, we can let $\phi_j = K_{\max}^{-1/2}$ (although more elaborate choices for ϕ_j could be considered; see for example [8]). Then the factor $\phi_j/\sqrt{2}$ in (21) can be absorbed along with the k in (22).

One important difficulty with (21) is that $\sigma_{\min}(H_k)$ is required to pick η_j at the start of step j, but H_k is not available until the final step k. A similar problem occurs in GMRES with inexact matrix-vector products; see [19, 6] and the comments in Section 4. In our experience, is often possible to replace $\sigma_{\min}(H_k)$ in (21) by 1, without significantly affecting the convergence of GMRES. This leads to following:

$$\text{Aggressive threshhold :} \qquad \eta_j = \epsilon\frac{\|b\|_2}{\|t_{j-1}\|_2}, \quad j = 1, 2, \ldots. \tag{24}$$

In exact arithmetic, $\sigma_{\min}(H_k)$ is bounded below by $\sigma_{\min}(A)$. If the smallest singular value of A is known, one can estimate $\sigma_{\min}(H_k) \approx \sigma_{\min}(A)$ in (21), leading to the following:

$$\text{Conservative threshhold :} \qquad \eta_j = \epsilon\, \sigma_{\min}(A) \frac{\|b\|_2}{\|t_{j-1}\|_2}, \quad j = 1, 2, \ldots. \tag{25}$$

This prevents potential early stagnation of the residual norm, but is often unnecessarily stringent. (It goes without saying that if the conservative threshold is less than $u\|A\|_2$, where u is the machine precision, then the criterion is vacuous: according to this criterion no inexact inner products can be carried out at iteration j.) Numerical examples are given in Section 5.

4 Incorporating inexact matrix-vector products

As mentioned in the introduction, there is already a literature on the use of inexact matrix-vector products in GMRES. These results are obtained by assuming that the Arnoldi vectors are orthonormal and analyzing the inexact Arnoldi relation

$$AV_k + E_k = V_{k+1} H_k, \qquad V_k^T V_k = I.$$

In practice, however, the computed Arnoldi vectors are very far from being orthonormal, even when all computations are performed in double precision arithmetic; see for example [5, 9, 14].

The purpose of this section is to show that the framework used in [19] and [6] to analyze inexact matrix-vector products in GMRES is still valid when the orthogonality of the Arnoldi vectors is lost, i.e., under the inexact Arnoldi relation

$$AV_k + E_k = V_{k+1} H_k, \qquad V_k^T V_k = I - F_k. \tag{26}$$

This settles a question left open in [19, Section 6].

Throughout we assume that $\|F_k\|_2 \leq \delta < 1$. Then from Theorem 1 there exists a symmetric positive definite matrix $W = I_n + M \in \mathbb{R}^{n\times n}$ such that $V_{k+1}^T W V_{k+1} = I_{k+1}$, and with singular values bounded as in (9).

4.1 Bounding the residual gap

As in previous sections, we use $x_k = V_k y_k$ to denote the computed GMRES iterate, with $r_k = b - Ax_k$ for the actual residual vector and $t_k = \beta_1 e_1 - H_k y_k$ for the residual vector updated in the GMRES iterations. From

$$\|r_k\|_2 \leq \|r_k - V_{k+1} t_k\|_2 + \|V_{k+1} t_k\|_2,$$

if

$$\max\{\, \|r_k - V_{k+1} t_k\|_2,\, \|V_{k+1} t_k\|_2 \,\} \leq \frac{\epsilon}{2} \|b\|_2 \tag{27}$$

then

$$\|r_k\|_2 \leq \epsilon \|b\|_2. \tag{28}$$

From the fact that the columns of $W^{1/2} V_{k+1}$ are orthonormal as well as (9), we obtain

$$\|V_{k+1} t_k\|_2 \leq \|W^{-1/2}\|_2 \|W^{1/2} V_{k+1} t_k\|_2 = \|W\|_2^{-1/2} \|t_k\|_2 \leq \sqrt{1+\delta}\, \|t_k\|_2.$$

In GMRES, $\|t_k\|_2 \to 0$ with increasing k, which implies that $\|V_{k+1}t_k\|_2 \to 0$ as well. Therefore, we focus on bounding the residual gap $\|r_k - V_{k+1}t_k\|_2$ in order to satisfy (27) and (28).

Suppose the matrix-vector products in the Arnoldi algorithm are computed inexactly, i.e., line 4 in Algorithm 1 is replaced by

$$w_j = (A + \mathcal{E}_j)v_j, \tag{29}$$

where $\|\mathcal{E}_j\|_2 \leq \epsilon_j$ for some given tolerance ϵ_j. Then in (26),

$$E_k = \left[\mathcal{E}_1 v_1, \mathcal{E}_2 v_2, \ldots, \mathcal{E}_k v_k\right]. \tag{30}$$

The following theorem bounds the residual gap at step k in terms of the tolerances δ and ϵ_j, for $j = 1, \ldots, k$.

Theorem 4. *Suppose that the inexact Arnoldi relation* (26) *holds, where E_k is given in* (30) *with $\|\mathcal{E}_j\|_2 \leq \epsilon_j$ for $j = 1, \ldots, k$, and $\|F_k\|_2 \leq \delta < 1$. Then the resulting residual gap satisfies*

$$\|r_k - V_{k+1}t_k\|_2 \leq \|H_k^\dagger\|_2 \sum_{j=1}^{k} \epsilon_j \|t_{j-1}\|_2. \tag{31}$$

Proof. From (26) and (30),

$$\begin{aligned}
\|r_k - V_{k+1}t_k\|_2 &= \|b - AV_k y_k - V_{k+1}t_k\|_2 \\
&= \|b - V_{k+1}H_k y_k + E_k y_k - V_{k+1}t_k\|_2 \\
&= \|V_{k+1}(\beta_1 e_1 - H_k y_k) + E_k y_k - V_{k+1}t_k\|_2 \\
&= \|E_k y_k\|_2 = \left\| \sum_{j=1}^{k} \mathcal{E}_j v_j e_j^T y_k \right\|_2 \leq \sum_{j=1}^{k} \epsilon_j |e_j^T y_k|.
\end{aligned}$$

The result then follows from (18). □

4.2 A strategy for picking the ϵ_j

Theorem 4 suggests the following strategy for picking the tolerances ϵ_j that bound the level of inexactness $\|\mathcal{E}_j\|_2$ in the matrix-vector products in (29). Similarly to Theorem 3, let ϕ_j be any positive numbers such that $\sum_{j=1}^{k} \phi_j = 1$. If for all steps $j = 1, \ldots, k$,

$$\epsilon_j \leq \frac{\phi_j \epsilon \sigma_{\min}(H_k)}{2} \frac{\|b\|_2}{\|t_{j-1}\|_2}, \tag{32}$$

then from (31) the residual gap in (27) satisfies

$$\|r_k - V_{k+1}t_k\|_2 \leq \frac{\epsilon}{2}\|b\|_2.$$

Interestingly, this result is independent of δ. Similarly to (21), the criterion for picking ϵ_j at step j involves H_k that is only available at the final step k. A large number of numerical experiments [6, 3] indicate that $\sigma_{\min}(H_k)$ can often be replaced by 1. Absorbing the factor $\phi_j/2$ into ϵ in (32) and replacing $\sigma_{\min}(H_k)$ by 1 or by $\sigma_{\min}(A)$ leads, respectively, to the same aggressive and conservative thresholds for ϵ_j as we obtained for η_j in (24) and in (25). This suggests that matrix-vector products and inner products in GMRES can be computed with the same level of inexactness. We illustrate this with numerical examples in the next section.

5 Numerical examples

We illustrate our results with a few numerical examples. We run GMRES with different matrices A and right-hand sides b, and compute the inner products and matrix-vector products inexactly as in (10) and (29). We pick η_{ij} randomly, uniformly distributed between $-\eta_j$ and η_j, and pick $\mathcal{E}_j$ to be a matrix of independent standard normal random variables, scaled to have norm ϵ_j. Thus we have

$$|\eta_{ij}| \leq \eta_j, \qquad \|\mathcal{E}_j\|_2 \leq \epsilon_j,$$

for chosen tolerances η_j and ϵ_j. Throughout we use the same level of inexactness for inner products and matrix-vector products, i.e., $\eta_j = \epsilon_j$.

In our first example, A is the 100×100 Grcar matrix of order 5. This is a highly non-normal Toeplitz matrix. The right hand side is $b = A[\sin(1), \ldots, \sin(100)]^T$. Results are shown in Figure 1. The solid green curve is the relative residual $\|b - Ax_k\|_2/\|b\|_2$. For reference, the dashed blue curve is the relative residual if GMRES is run in double precision. The full magenta curve corresponds to the loss of orthogonality $\|F_k\|_2$ in (11). The black dotted curve is the chosen tolerance η_j.

In Example 1(a),

$$\eta_j = \epsilon_j = \begin{cases} 10^{-8}\|A\|_2, & \text{for } 20 \leq j \leq 30, \\ 10^{-4}\|A\|_2, & \text{for } 40 \leq j \leq 50, \\ 2^{-52}\|A\|_2, & \text{otherwise.} \end{cases}$$

The large increase in the inexactness of the inner products at iterations 20 and 40 immediately leads to a large increase in $\|F_k\|_2$. This clearly illustrates the connection between the inexactness of the inner products and the loss of orthogonality in the Arnoldi vectors. As proven in Theorem 2, until $\|F_k\|_2 \approx 1$, the residual norm is the same as it would have been had all computations been performed in double precision. Due to its large increases at iterations 20 and 40, $\|F_k\|_2$ approaches 1, and the residual norm starts to stagnate, long before a backward stable solution is obtained.

In Example 1(b), the tolerances are chosen according to the aggressive criterion (24) with $\epsilon = 2^{-52}\|A\|_2$. With this judicious choice, $\|F_k\|_2$ does not reach 1, and the residual norm does not stagnate, until a backward stable solution is obtained.

In our second example, A is the matrix 494_bus from the SuiteSparse matrix collection [4]. This is a 494×494 matrix with condition number $\kappa_2(A) \approx 10^6$. The right hand side is once again $b = A[\sin(1), \ldots, \sin(100)]^T$.

Results are shown in Figure 2. In Example 2(a), tolerances are chosen according to the aggressive threshhold (24) with $\epsilon = 2^{-52}\|A\|_2$. In this more ill-conditioned problem, the residual norm starts to stagnate before a backward stable solution is obtained. In Example 2(b), the tolerances are chosen according to the conservative threshhold (25) with $\epsilon = 2^{-52}\|A\|_2$, and there is no more such stagnation. Because of these lower tolerances, the inner products and matrix-vector products have to be performed in double precision until about iteration 200. This example illustrates the tradeoff between the level of inexactness and the maximum attainable accuracy. If one requires a backward stable solution, the more ill-conditioned the matrix A is, the less opportunity there is for performing floating-point operations inexactly in GMRES.

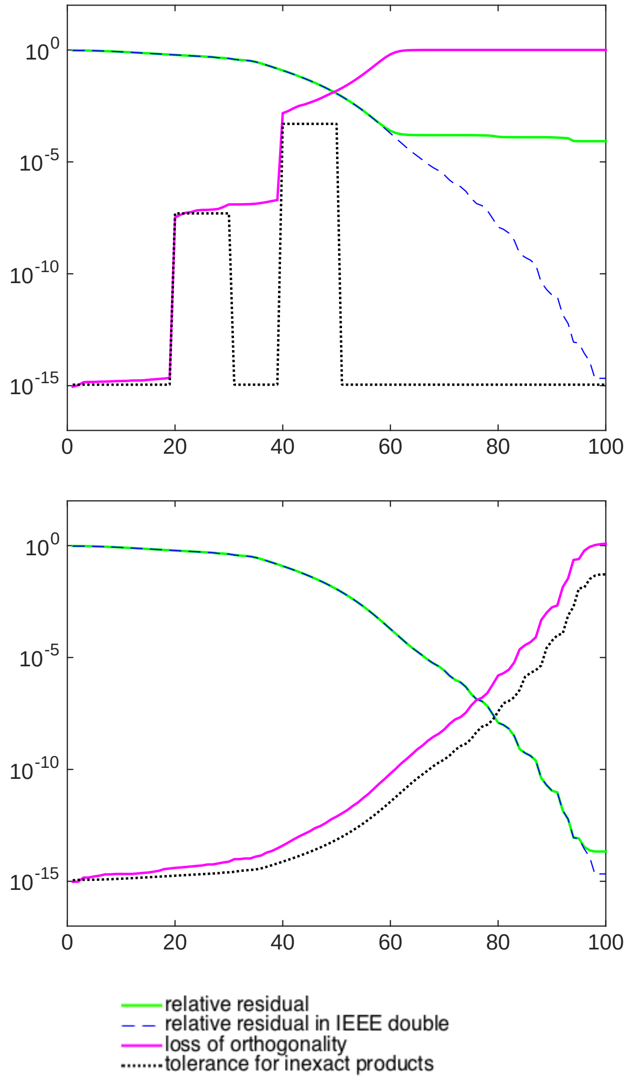

Figure 1: GMRES in variable precision: Grcar matrix.

6 Conclusion

We have shown how inner products can be performed inexactly in MGS-GMRES without affecting the convergence or final achievable accuracy of the algorithm. We have also shown that a known framework for inexact matrix-vector products is still valid despite the loss of orthogonality in the Arnoldi vectors. It would be interesting to investigate the impact of scaling or preconditioning on these results. Additionally,

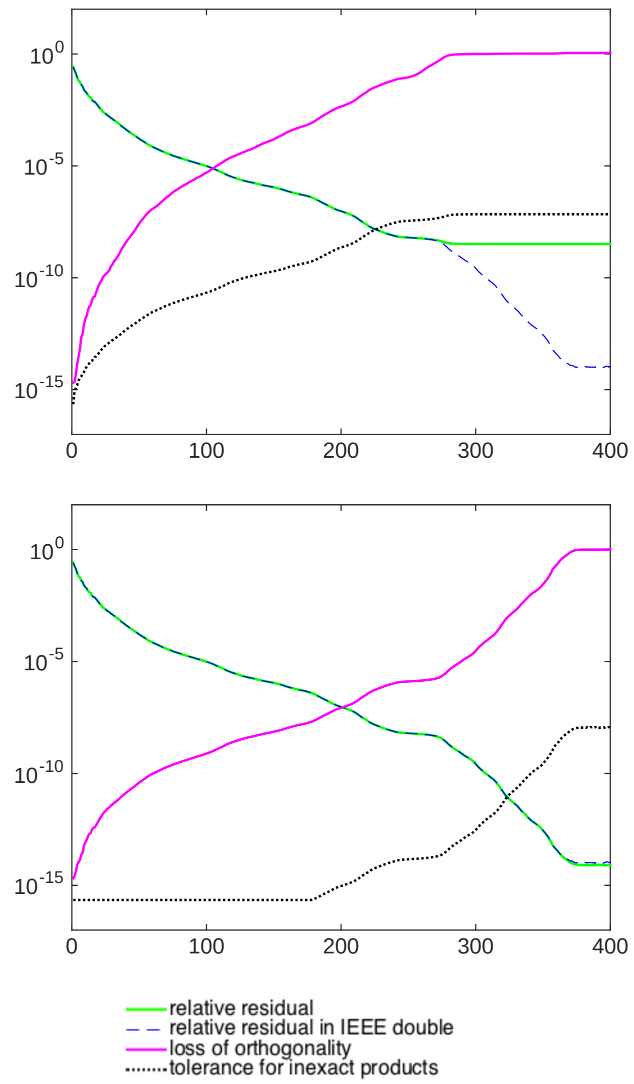

Figure 2: GMRES in variable precision: 494_bus matrix

in future work, we plan to address the question of how much computational savings can be achieved by this approach on massively parallel computer architectures.

Appendix

A Proof of (14)

In line 7 of Algorithm 1, in the ℓth pass of the inner loop at step j, we have

$$w_j^{(\ell)} = w_j^{(\ell-1)} - h_{\ell j} v_\ell \tag{A1}$$

for $\ell = 1, \ldots, j$ and with $w_j^{(0)} = Av_j$. Writing this equation for $\ell = i+1$ to j, we have

$$\begin{aligned} w_j^{(i+1)} &= w_j^{(i)} - h_{i+1,j} v_{i+1}, \\ w_j^{(i+2)} &= w_j^{(i+1)} - h_{i+2,j} v_{i+2}, \\ &\vdots \\ w_j^{(j)} &= w_j^{(j-1)} - h_{j,j} v_j. \end{aligned}$$

Summing the above and cancelling identical terms that appear on the left and right hand sides gives

$$w_j^{(j)} = w_j^{(i)} - \sum_{\ell=i+1}^{j} h_{\ell j} v_\ell.$$

Because $w_j^{(j)} = v_{j+1} h_{j+1,j}$, this reduces to

$$w_j^{(i)} = \sum_{\ell=i+1}^{j+1} h_{\ell j} v_\ell. \tag{A2}$$

Because the inner products h_{ij} are computed inexactly as in (10), from (A1) we have

$$\begin{aligned} w_j^{(i)} &= w_j^{(i-1)} - h_{ij} v_i \\ &= w_j^{(i-1)} - (v_i^T w_j^{(i-1)} + \eta_{ij}) v_i \\ &= (I - v_i v_i^T) w_j^{(i-1)} - \eta_{ij} v_i. \end{aligned}$$

Therefore,

$$v_i^T w_j^{(i)} = -\eta_{ij}.$$

Multiplying (A2) on the left by $-v_i^T$ gives

$$\eta_{ij} = -\sum_{\ell=i+1}^{j+1} h_{\ell j} (v_i^T v_\ell), \tag{A3}$$

which is the entry in position (i, j) of the matrix equation

$$\begin{bmatrix} \eta_{11} & \ldots & \eta_{1k} \\ & \ddots & \vdots \\ & & \eta_{kk} \end{bmatrix} = - \begin{bmatrix} v_1^T v_2 & \ldots & v_1^T v_{k+1} \\ & \ddots & \vdots \\ & & v_k^T v_{k+1} \end{bmatrix} \begin{bmatrix} h_{21} & \ldots & h_{2k} \\ & \ddots & \vdots \\ & & h_{k+1,k} \end{bmatrix},$$

i.e., (14).

B Proof of Lemma 2

For any $\gamma > 0$, the smallest singular value of the matrix $\left[\beta\gamma e_1, H_k D_k^{-1}\right]$ is the scaled total least squares (STLS) distance [16] for the estimation problem $H_k D_k^{-1} z \approx \beta e_1$. As shown in [15], it can be bounded by the least squares distance

$$\min_z \|\beta e_1 - H_k D_k^{-1} z\|_2 = \|\beta e_1 - H_k D_k^{-1} z_k\|_2 = \|\beta e_1 - H_k y_k\|_2 = \|t_k\|_2,$$

where $z_k = D_k y_k$. From [15, Theorem 4.1], we have

$$\frac{\|t_k\|_2}{\left(\gamma^{-2} + \|D_k y_k\|_2^2/(1-\tau_k^2)\right)^{1/2}} \le \sigma_{\min}\left(\left[\beta\gamma e_1, H_k D_k^{-1}\right]\right) \le \gamma\|t_k\|_2, \tag{B1}$$

provided $\tau_k < 1$, where

$$\tau_k \equiv \frac{\sigma_{\min}\left(\left[\beta\gamma e_1, H_k D_k^{-1}\right]\right)}{\sigma_{\min}\left(H_k D_k^{-1}\right)}.$$

We now show that if $\gamma = (\epsilon\|b\|_2)^{-1}$ and D_k satisfies (19), then $\tau_k \le 1/\sqrt{2}$. From the upper bound in (B1) we immediately have

$$\sigma_{\min}\left(\left[\beta\gamma e_1, H_k D_k^{-1}\right]\right) \le \gamma\|t_k\|_2 = \frac{\|t_k\|_2}{\epsilon\|b\|_2}.$$

Also,

$$\sigma_{\min}\left(H_k D_k^{-1}\right) = \min_{z\neq 0} \frac{\|H_k D_k^{-1} z\|_2}{\|z\|_2} = \min_{z\neq 0} \frac{\|H_k z\|_2}{\|D_k z\|_2} \ge \min_{z\neq 0} \frac{\|H_k z\|_2}{\|D_k\|_2\|z\|_2} = \frac{\sigma_{\min}(H_k)}{\|D_k\|_2}.$$

Therefore, if (19) holds,

$$\tau_k \le \frac{\|t_k\|_2}{\epsilon\|b\|_2}\frac{\|D_k\|_2}{\sigma_{\min}(H_k)} \le \frac{1}{\sqrt{2}}.$$

Substituting $\gamma = (\epsilon\|b\|_2)^{-1}$ and $\tau_k \le 1/\sqrt{2}$ into (B1) gives (20).

References

[1] A. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT Numerical Mathematics, 7 (1967), pp. 1–21.

[2] A. Björck and C. Paige, *Loss and recapture of orthogonality in the Modified Gram-Schmidt algorithm*, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 176–190.

[3] A. Bouras and V. Frayssé, *Inexact matrix-vector products in Krylov methods for solving linear systems: A relaxation strategy*, SIAM Journal on Matrix Analysis and Applications, 26 (2006), pp. 660–678.

[4] T. A. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM Transactions on Mathematical Software, 38 (2011), pp. 1–25.

[5] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of GMRES*, BIT Numerical Mathematics, 35 (1995), pp. 309–330.

[6] J. V. D. Eshof and G. Sleijpen, *Inexact Krylov subspace methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 26 (2004), pp. 125–153.

[7] L. Giraud, S. Gratton, and J. Langou, *Convergence in backward error of relaxed GMRES*, SIAM Journal on Scientific Computing, 29 (2007), pp. 710–728.

[8] S. Gratton, E. Simon, and P. Toint, *Minimizing convex quadratic with variable precision Krylov methods*, arXiv, abs/1807.07476 (2018).

[9] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behaviour of the Modified Gram-Schmidt GMRES implementation*, BIT Numerical Mathematics, 37 (1997), pp. 706–719.

[10] S. Güttel and J. Pestana, *Some observations on weighted GMRES*, Numerical Algorithms, 67 (2014), pp. 733–752.

[11] N. J. Higham, *A multiprecision world*, SIAM News, 50 (2017).

[12] S. J. Leon, A. Björck, and W. Gander, *Gram-Schmidt orthogonalization: 100 years and more*, Numerical Linear Algebra with Applications, 20 (2013), pp. 492–532.

[13] C. Paige, *A useful form of unitary matrix obtained from any sequence of unit 2-norm n-vectors*, SIAM Journal on Matrix Analysis and Applications, 31 (2009), pp. 565–583.

[14] C. Paige, M. Rozložník, and Z. Strakoš, *Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM Journal on Matrix Analysis and Applications, 28 (2006), pp. 264–284.

[15] C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numerische Mathematik, 91 (2002), pp. 93–115.

[16] C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numerische Mathematik, 91 (2002), pp. 117–146.

[17] C. Paige and W. Wülling, *Properties of a unitary matrix obtained from a sequence of normalized vectors*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 526–545.

[18] J. Pestana and A. J. Wathen, *On the choice of preconditioner for minimum residual methods for non-Hermitian matrices*, Journal of Computational and Applied Mathematics, 249 (2013), pp. 57 – 68.

[19] V. Simoncini and D. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM Journal on Scientific Computing, 25 (2003), pp. 454–477.